Available online http://ccforum.com/content/13/1/201

Review

Bench-to-bedside review: The importance of the precision of the reference technique in method comparison studies – with specific reference to the measurement of cardiac output

Maurizio Cecconi[1,2], Andrew Rhodes[2], Jan Poloniecki[3], Giorgio Della Rocca[1] and R Michael Grounds[2]

[1]Department of Anesthesia and Intensive Care, Azienda Ospedaliero Universitaria Udine, Piazzale Santa Maria della Misericordia, 33100 Udine, Italy
[2]Department of Intensive Care Medicine, St George's Hospital, London, SW17 0QT, UK
[3]Community Health Sciences, St George's, University of London, SW17 0RE, UK

Corresponding author: Maurizio Cecconi, mauriziocecconi@hotmail.com

Published: 13 January 2009
This article is online at http://ccforum.com/content/13/1/201

Critical Care 2009, **13:**201 (doi:10.1186/cc7129)

Abstract

Bland-Altman analysis is used for assessing agreement between two measurements of the same clinical variable. In the field of cardiac output monitoring, its results, in terms of bias and limits of agreement, are often difficult to interpret, leading clinicians to use a cutoff of 30% in the percentage error in order to decide whether a new technique may be considered a good alternative. This percentage error of ± 30% arises from the assumption that the commonly used reference technique, intermittent thermodilution, has a precision of ± 20% or less. The combination of two precisions of ± 20% equates to a total error of ± 28.3%, which is commonly rounded up to ± 30%. Thus, finding a percentage error of less than ± 30% should equate to the new tested technique having an error similar to the reference, which therefore should be acceptable. In a worked example in this paper, we discuss the limitations of this approach, in particular in regard to the situation in which the reference technique may be either more or less precise than would normally be expected. This can lead to inappropriate conclusions being drawn from data acquired in validation studies of new monitoring technologies. We conclude that it is not acceptable to present comparison studies quoting percentage error as an acceptability criteria without reporting the precision of the reference technique.

Introduction

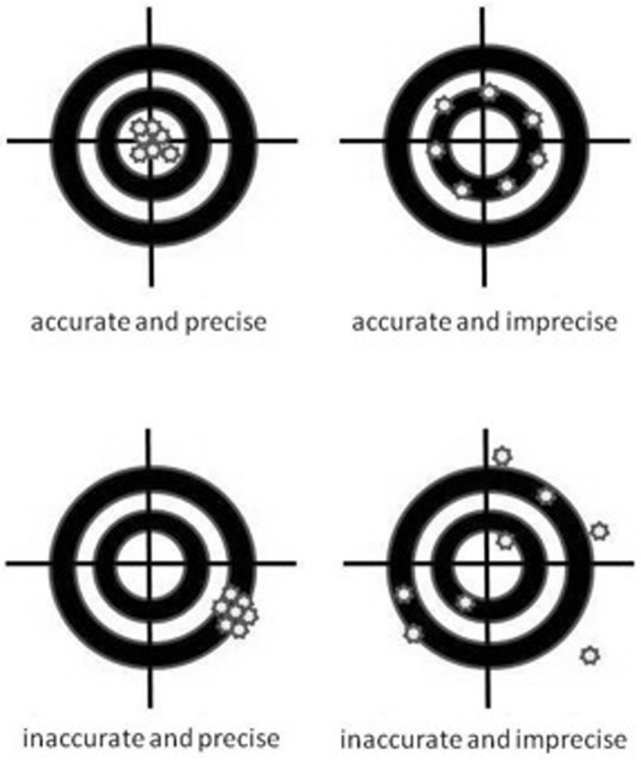
In 1986, Bland and Altman [1] first suggested their statistical method for assessing agreement between two measurements of the same clinical variable. They described the 'Bland-Altman' plot as a mechanism for displaying and describing data from studies in which one variable is measured by two different techniques. Since then, this 'plot' together with the associated analysis has become the recognised statistical methodology for studies validating new measuring or monitoring tools against a reference technique [2,3]. The Bland-Altman plot is able to provide researchers with a graphical representation of their data and also a number of objective measures of how well the data series agree with each other:

1. The bias: the average of all the differences.
2. The standard deviation around the bias.
3. The limits of agreement: the limits within which 95% of all the points fall on either side of the bias (that is, ± 1.96 times the standard deviation around the bias).

These variables can be used to describe the accuracy and precision of any given device. The accuracy describes how close to the actual or real value the measurement is, whereas the precision describes how close the values of repeated measurements are. A good method should be both accurate and precise. A visual example may clarify this point (Figure 1). If we imagine a cardiac output monitor as a gun that is used to shoot a target (the cardiac output), we can classify accuracy as the characteristic of being able to shoot close to the centre of the bull's-eye. Precision is related to how close repeated shots are to each other. How can we use these concepts when looking at the Bland-Altman plot? First of all, we have to imagine that our reference technique is very reliable. Otherwise, the effect would be that of a 'moving target'. If the bias then is low, it means that the accuracy is high. Limits of agreement refer to how precise the measurements are. So if they are narrow, the precision is high; if they are large, the precision is low. The bias therefore allows an estimate to be made of the accuracy of the new device, and the limits of agreement allow an estimate of the precision or random error around the bias. An ideal result therefore would have a very small bias with tight limits of agreement. These

CE = coefficient of error; CV = coefficient of variation; ITD = intermittent thermodilution; LSC = least significant change; OD = oesophageal Doppler; PE = percentage error.

Figure 1



Bull's-eye representation of accuracy and precision. With respect to the Bland-Altman plot, accurate measurements mean small bias and precise measurements mean narrow limits of agreement.

descriptive terms are commonly used both to describe the results of studies and to justify the conclusions. There is no real consensus, however, in how these statistical terms relate to any given variable and this has led to much confusion in how to interpret studies and therefore in whether (or not) to accept new measuring or monitoring devices into routine clinical practice.

Validation of cardiac output monitoring devices

Ideally, any reference technique used should be able to provide an accurate and precise measurement of cardiac output. However, in clinical practice or human research, this is rarely possible. The ideal reference method of measuring cardiac output has not been described. However, the most commonly used reference technique is an averaged set of thermodilution curves taken from a pulmonary artery catheter. This technique has been well studied and the level of precision, if properly performed, is understood. In recent years, there have been a large number of studies published in which a new method of measuring cardiac output has been assessed using intermittent thermodilution (ITD) from the pulmonary artery catheter as the reference technique [4-10]. All of these studies have used the Bland-Altman methodology to describe their data. In most studies, the results have demonstrated a small bias but relatively wide limits of agreement. For instance, Sander and colleagues [11] demonstrated that, in comparison with ITD, the Vigileo/Flotrac device (Edwards Lifesciences LLC, Irvine, CA, USA) had a bias of 0.6 litres per minute and limits of agreement of between −2.2 to +3.4 litres per minute. These results were reported as demonstrating that the new tested device, the Vigileo, was not a good measure of cardiac output compared with the reference technique. However, it is not clear from this paper, like many others reported before [12-14], what would have been acceptable limits of agreement in order for the study to confirm the efficacy of the new tool. To allow a conclusion to be drawn from the data, the authors should have made an *a priori* description of what they perceived to be acceptable limits of agreement. Unless this is described before the study is commenced, it becomes very difficult to make sensible conclusions from the data.

To understand how wide the limits of agreement may be, it is important to understand that with the Bland-Altman plot it is possible to assess two independent methods of measuring the same variable, each of which has its own inherent error. The limits of agreement describe the variance around the bias, which is in itself an averaged value taken from each pair of study measurements. The limits of agreement also relate to the population being studied. For instance, if the limits of agreement are ± 1 litres per minute, this would be good for a hyperdynamic population of patients with a mean cardiac output of 10 litres per minute, but not so good for a paediatric population with a mean cardiac output of 2 litres per minute. Critchley and Critchley [15], in their meta-analysis of cardiac output validation studies, suggested a solution to this problem. They proposed that the percentage error (PE) of the limits of agreement, as compared with the population mean, be used to describe the agreement and that this could be used as a cutoff for whether to accept a new technique [15]. The basis of this approach is that, in order to accept the new technology (unless it heralds other significant advantages), the level of accuracy and precision should at the very least equal that of the reference technique. In statistical terms, the random error that produces imprecision from a single measurement is described by the coefficient of variation (CV). This is calculated as the standard deviation divided by the mean. When more than one measurement is used to produce the overall result (for instance, when averaging three thermodilution curves), the coefficient of error (CE), as calculated from the following equation, is more appropriate:

$$CE = CV / \sqrt{n} \quad (1)$$

where CE = coefficient of variation of average of n measurements, CV = coefficient of variation of single measurements, and n = number of repeated measurements.

When one only measurement is used, the CE is equal to the CV. The precision of the technique is considered to be two

times the CV or two times the CE. From now on, we will refer to 2CV or to 2CE as precision. Critchley and Critchley [15] looked at studies assessing oesophageal Doppler (OD) ultrasound techniques as a measurement of cardiac output. They compared these against ITD cardiac output from the pulmonary artery catheter, which they described as having a precision of ± 20%. They suggested that, in order for the new device (Doppler) to be accepted, it should have an equivalent precision (that is, 20%). Therefore, the PE from the Bland-Altman plot, taken from the following equation, should be less than 28.3% [15]:

$$CV_{a-b} = \sqrt{[(CV_a)^2 + (CV_b)^2]} \quad (2)$$

where CV_{a-b} = CV of the differences between the two methods, CV_a = CV of method a, and CV_b = CV of method b.

This has been simplified by many authors to be a requirement that a new technology have a PE from the Bland-Altman plot of less than ± 30% [10,15-17]. In our opinion, it is quite clear that this ± 30% margin for the PE hides some important information and, if used without understanding the background behind it, may lead to erroneous conclusions being drawn from study results. The 30% limit is contributed to by two separate levels of precision, which when combined add up to this value of ± 30% error. It should be intuitive, therefore, to understand that the precision of the reference technique is extremely important when assessing the combined error of the two. This has been studied extensively with ITD and the variance can range from 5% to 15% depending on the technique used. The main limitation of this ± 30% cutoff, therefore, is that it relies on the fact that the precision of ITD is always the same and is usually around ± 20%. If the reference technique is performed with a high degree of rigour, its precision may actually be significantly less than the 20% allowed for in the above equation. This may lead to the acceptance of a studied technique with an inappropriate level of precision. It is obvious that there is a relationship between the two individual errors and the combined sum (Figure 2).

If:
Precision for method a, $precision_a$, $2 \times CV_a$
Precision for method b, $precision_b$, $2 \times CV_b$
Percentage error is $PE_{a-b} = 2CV_{a-b}$

Then:

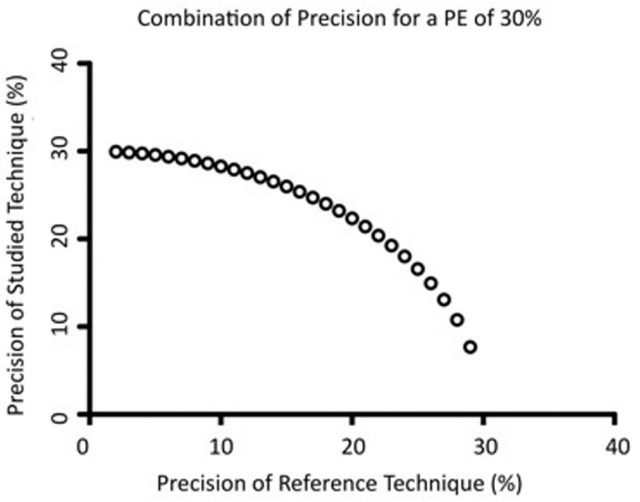
$$PE_{a-b} = \sqrt{[(precision_a)^2 + (precision_b)^2]} \quad (3)$$

If:
PE_{a-b} from the Bland-Altman plot is known and $precision_a$ is known,

Then:

$$Precision_b = \sqrt{[(PE_{a-b})^2 - (precision_a)^2]}$$

Figure 2



Different combinations of precision for a reference and a new method that can lead to a percentage error (PE) of 30%. A 30% PE can derive from several combinations of precisions for the two methods compared.

Therefore, we would suggest that, in any study in which a new technique is to be validated against a reference, the precision of the reference technique within the study be measured and quoted, thus enabling an estimation of the new technique to be made. Then whatever reference technique is used in studies assessing a new cardiac output monitor, there should always be a description of the error of that technique as obtained within the study. These concepts hold true for any study assessing a new methodology of measurement against a reference in clinical science.

Worked example

Table 1 describes data taken from two independent measures of cardiac output (A and B). The average cardiac outputs by the reference technique and test technique were 8.0 and 8.2 litres per minute, respectively. The average of these was 8.1 litres per minute. In this example, measurements were taken at times of stable haemodynamic situations and the reference technique was ITD from a pulmonary artery catheter measured from four independent and averaged curves. The standard Bland-Altman plot is described in Figure 3. The bias between the two techniques is 0.2 litres per minute with limits of agreement around the bias of ± 2.5 litres per minute. This provides a PE for the agreement between the two techniques of ± 30%. At first glance, this would suggest that the new technique almost fulfills the criteria to be within a ± 30% error rate. If the monitor has other advantages (perhaps being less invasive, cheaper, and easier to set up), this may be considered adequate for normal practice. However, to understand the precision of the new technique, it is necessary to look more carefully at the

Table 1

Cardiac output in 20 patients: repeated measurements with the reference technique and single test measurements

Patient	CO1, L/min	CO2, L/min	CO3, L/min	CO4, L/min	Mean CO, L/min	CV, ± %	CE, ± %	Studied CO, L/min
1	6.5	8.8	7.0	7.7	7.5	13	7	6.9
2	11.8	15.1	14.3	12.8	13.5	11	5	10.9
3	7.8	6.7	6.5	6.6	6.9	9	4	6.7
4	7.0	7.3	6.5	7.1	7.0	5	2	8.2
5	6.1	6.7	7.6	6.5	6.7	9	5	5.9
6	13.2	14.4	12.8	13.6	13.5	5	3	14.2
7	13.1	11.7	14.7	13.3	13.2	9	5	11.9
8	6.1	6.3	6.6	7.4	6.6	9	4	7.3
9	16.2	12.7	13.3	14.4	14.2	11	5	13.8
10	5.2	5.2	3.9	5.6	5.0	15	7	6.0
11	6.8	7.4	6.1	6.4	6.7	8	4	7.1
12	8.2	7.7	8.2	7.6	7.9	4	2	7.4
13	6.7	5.5	6.9	5.8	6.2	11	5	6.8
14	3.9	4.3	4.7	5.0	4.5	11	5	3.9
15	6.1	6.6	6.7	4.9	6.1	14	7	7.8
16	8.0	8.5	8.0	8.2	8.2	3	1	10.2
17	8.0	6.9	7.5	8.1	7.6	7	4	9.4
18	7.0	6.3	7.5	6.5	6.8	8	4	5.6
19	8.2	7.5	8.8	8.3	8.2	7	3	10.1
20	4.5	3.9	4.1	4.6	4.3	8	4	4.5
Average	8.0	8.0	8.1	8.0	8.0	9	4	8.2

Mean cardiac output (CO) is the mean of the four measurements. CV is the coefficient of variation for a single measurement, and CE is the coefficient of error when four measurements are averaged. The studied CO is the single measurement for the studied technique.

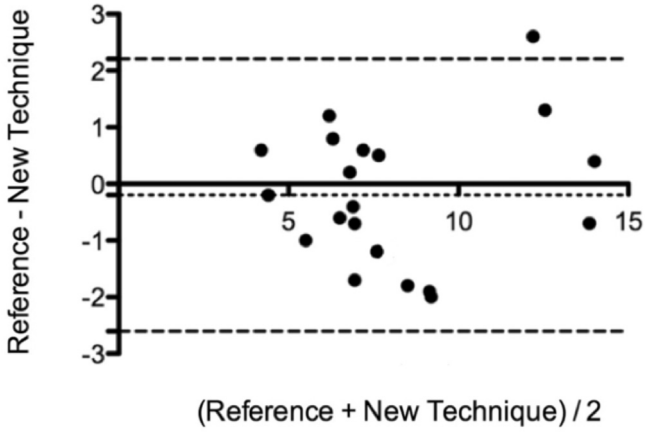
precision of the reference. In this example, as technique A was ITD, four measurement curves were performed enabling the CE of this technique under the study conditions to be calculated: 4% for four averaged curves. By using equation 2 (as described above), it is then possible to calculate the CV of the tested device, which in this case is 15%. It is then obvious that, although the combined PE is almost adequate, the precision of the new technique is more than three times worse than the reference that it is attempting to replace.

For the purposes of this example, it is helpful to envision the situation of the reference technique (ITD) being performed at a number of differing levels of precision. For example, if the comparison is done with one curve with a CV of 9%, then for a studied technique with an error of 15% the PE from the Bland-Altman plot is 34%, which according to the Critchley and Critchley criteria is not acceptable (Table 2). On the other hand, if the reference technique uses an average of four curves (CE of 4%), then for the same technique as before (error of 15%) the PE for the Bland-Altman plot is ± 30%, which according to the Critchley and Critchley criteria would be acceptable (Table 2 and Figure 4).

Clinical implications of understanding the error for a cardiac output monitoring device

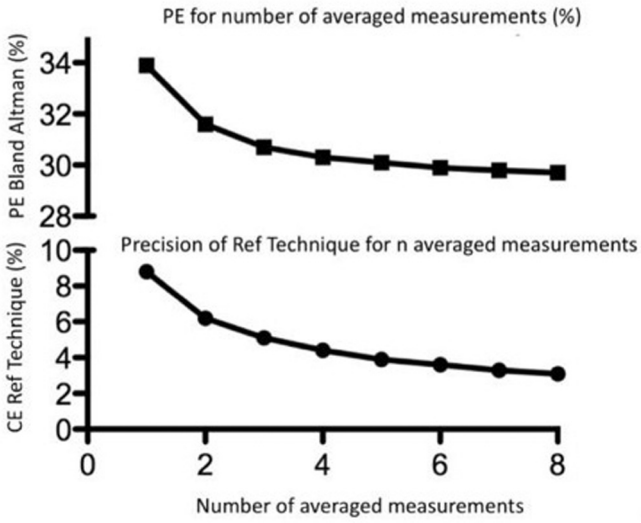
The understanding of how precise a monitor is allows us to appreciate two important concepts. The first relates (as discussed above) to how one monitor compares with another in terms of accuracy, and the second relates to how the monitor performs in normal clinical practice. If we assume that the CE for ITD in normal clinical practice is 10%, what does it tell us? For an individual patient, a CE of 10% implies that the exact value of cardiac output lies with 95% certainty somewhere in a band between ± 20% (two times the error) of the measured level. It is especially important to understand the precision of these new tools when using them to target fixed resuscitation endpoints (for instance, perioperative haemodynamic optimisation protocols that aim to target an absolute value of oxygen delivery index of 600 mL/min*m^2 [18,19]). An error of 15% would mean that the measured cardiac output

Figure 3



Bland-Altman plot for new technique versus reference technique. Dotted lines represent bias and limits of agreement. Data from Table 1 are used.

Figure 4



Precision of the reference technique for n averaged measurements and the corresponding percentage error (PE) from the Bland-Altman plot for a fixed level of precision of the studied technique (29%). The PE can change simply by using a more or less precise reference technique, even when the precision of the studied technique is not changed. This may lead to the acceptance of a studied technique even though its performance in terms of precision stays the same. CE, coefficient of error.

of 4.5 litres per minute could be anything from 3 to 6 litres per minute (95% confidence). This may have profound clinical implications.

In many clinical situations, there is no 'normal' cardiac output for any individual patient at any specific time point. Most clinicians, therefore, use these devices to see how the physiology of the patient changes following an intervention. A standard technique would be to perform a fluid challenge with the aim of increasing the cardiac output by 10% from the baseline value. It is obvious that, in order for a monitor to be used to detect this 10% change, it must have a level of precision that can detect this change and this is traditionally done with 95% certainty. Measuring a change, however, does not necessarily mean that the physiological status of the patient has changed. The error of the measuring technique is directly related to the magnitude of the least significant change (LSC). The LSC is the minimum change that needs to

Table 2

Effect of the number of measurements of the reference technique on the percentage error

Measurements for the reference technique, number	Ref precision, ± percentage	Study precision, ± percentage	PE, ± percentage	± 30% fulfilled
1	18	29	34	No
2	12	29	32	No
3	10	29	31	No
4	9	29	30	Yes
5	8	29	30	Yes
6	7	29	30	Yes
7	7	29	30	Yes
8	6	29	30	Yes

'Measurements for the reference technique' means the number of measurements averaged for the reference technique. 'Ref precision' is the precision for the reference technique according to the number of averaged measurements, 'study precision' is the precision of the studied technique as measured by the worked example, and PE is the percentage error for the Bland-Altman plot for the reference technique minus the studied technique. The '± 30% fulfilled' column shows whether the PE would be accepted according to a cutoff of 30%.

be measured by a device in order to recognise a real change and can be described by the following equation:

$$LSC = \text{precision} \sqrt{2}.$$

This means that the usually accepted 10% CE for ITD would allow measured changes to be trusted as real only if greater than 28.3%. Understanding the error in single patients, therefore, will give us an estimate in the single patient of whether a change has actually happened. Roeck and colleagues [20] measured stroke volume before and after a fluid challenge with ITD and with OD measured by two independent observers. There was a significant difference between the two observers measuring the same change (if any happened at all) and also between changes measured by the two techniques. In their study, the error for ITD was 8% (clinically acceptable) but, interestingly, was too high to consider measured changes of less than 22% in magnitude [20]. This may explain why the variation with the OD before and after the fluid challenge was higher than the ones recorded by ITD. As the authors stated, they found a higher-than-expected variability in the Doppler. This was to be expected from the variability in the reference technique.

Recommendations for validation studies of new cardiac output monitors

1. The reference technique should be as accurate and precise as possible.
2. The precision of the reference technique should be measured within the study.
3. The desired precision of the new technique should be described *a priori*.
4. The bias and limits of agreement between the two techniques should be quoted.
5. The precision of the new tested technique should be calculated.

Conclusions

As new technologies come into the marketplace, the requirement for validation studies will increase. To make a fair and valid comparison between new tools and more traditional 'gold standard' reference techniques, it is necessary to have a robust and sensitive mechanism for performing the studies and analysing the data. The understanding of the precision of a new device is vital prior to accepting it into clinical practice and prior to using it for significant therapeutic interventions. Therefore, measuring the error of the studied techniques should always be performed when comparing two methods. This approach can be used for any method comparison provided that the variance within the individuals of at least one of the methods can be estimated.

Competing interests

AR has received lecturing fees from Edwards Lifesciences LLC (Irvine, CA, USA) and LiDCO (Sawston, Cambridge, UK). The other authors declare that they have no competing interests.

References

1. Bland JM, Altman DG: **Statistical methods for assessing agreement between two methods of clinical measurement.** *Lancet* 1986, **1:**307-310.
2. Bland JM, Altman DG: **Comparing methods of measurement: why plotting difference against standard method is misleading.** *Lancet* 1995, **346:**1085-1087.
3. Bland JM, Altman DG: **Measuring agreement in method comparison studies.** *Stat Methods Med Res* 1999, **8:**135-160.
4. Hamilton TT, Huber LM, Jessen ME: **PulseCO: a less-invasive method to monitor cardiac output from arterial pressure after cardiac surgery.** *Ann Thorac Surg* 2002, **74:**S1408-1412.
5. Lichtenthal PR, Gordan D: **Testing the safety of Baxter continuous cardiac output monitoring system.** *J Clin Monit* 1996, **12:** 243-249.
6. Munro HM, Wood CE, Taylor BL, Smith GB: **Continuous invasive cardiac output monitoring—the Baxter/Edwards Critical-Care Swan Ganz IntelliCath and Viligance system.** *Clin Intensive Care* 1994, **5:**52-55.
7. Della Rocca G, Costa MG, Pompei L, Coccia C, Pietropaoli P: **Continuous and intermittent cardiac output measurement: pulmonary artery catheter versus aortic transpulmonary technique.** *Br J Anaesth* 2002, **88:**350-356.
8. Goedje O, Hoeke K, Lichtwarck-Aschoff M, Faltchauser A, Lamm P, Reichart B: **Continuous cardiac output by femoral arterial thermodilution calibrated pulse contour analysis: comparison with pulmonary arterial thermodilution.** *Crit Care Med* 1999, **27:**2407-2412.
9. Button D, Weibel L, Reuthebuch O, Genoni M, Zollinger A, Hofer CK: **Clinical evaluation of the FloTrac/Vigileo system and two established continuous cardiac output monitoring devices in patients undergoing cardiac surgery.** *Br J Anaesth* 2007, **99:** 329-336.
10. Mayer J, Boldt J, Schollhorn T, Rohm KD, Mengistu AM, Suttner S: **Semi-invasive monitoring of cardiac output by a new device using arterial pressure waveform analysis: a comparison with intermittent pulmonary artery thermodilution in patients undergoing cardiac surgery.** *Br J Anaesth* 2007, **98:**176-182.
11. Sander M, Spies CD, Grubitzsch H, Foer A, Muller M, von Heymann C: **Comparison of uncalibrated arterial waveform analysis in cardiac surgery patients with thermodilution cardiac output measurements.** *Crit Care* 2006, **10:**R164.
12. Barin E, Haryadi DG, Schookin SI, Westenskow DR, Zubenko VG, Beliaev KR, Morozov AA: **Evaluation of a thoracic bioimpedance cardiac output monitor during cardiac catheterization.** *Crit Care Med* 2000, **28:**698-702.
13. Neviere R, Mathieu D, Riou Y, Chagnon JL, Wattel F: **Non-invasive measurement of cardiac output in patients with acute lung injury using the carbon dioxide rebreathing method.** *Clin Intensive Care* 1994, **5:**172-175.
14. Jakobsen CJ, Melsen NC, Andresen EB: **Continuous cardiac output measurements in the perioperative period.** *Acta Anaesthesiol Scand* 1995, **39:**485-488.
15. Critchley LA, Critchley JA: **A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques.** *J Clin Monit Comput* 1999, **15:**85-91.
16. Pittman J, Bar-Yosef S, SumPing J, Sherwood M, Mark J: **Continuous cardiac output monitoring with pulse contour analysis: a comparison with lithium indicator dilution cardiac output measurement.** *Crit Care Med* 2005, **33:**2015-2021.
17. Bein B, Worthmann F, Tonner PH, Paris A, Steinfath M, Hedderich J, Scholz J: **Comparison of esophageal Doppler, pulse contour analysis, and real-time pulmonary artery thermodilution for the continuous measurement of cardiac output.** *J Cardiothorac Vasc Anesth* 2004, **18:**185-189.
18. Pearse R, Dawson D, Fawcett J, Rhodes A, Grounds RM, Bennett ED: **Early goal-directed therapy after major surgery reduces complications and duration of hospital stay. A randomised, controlled trial [ISRCTN38797445].** *Crit Care* 2005, **9:**R687.
19. Boyd O, Grounds RM, Bennett ED: **A randomized clinical trial of the effect of deliberate perioperative increase of oxygen delivery on mortality in high-risk surgical patients.** *JAMA* 1993, **270:**2699-2707.
20. Roeck M, Jakob SM, Boehlen T, Brander L, Knuesel R, Takala J: **Change in stroke volume in response to fluid challenge: assessment using esophageal Doppler.** *Intensive Care Med* 2003, **29:**1729-1735.